How a modern IR system operates

- (off-line)
  - building document representations and loading them into an internal index structure;
  - post-processing the retrieved documents
- (on-line)
  - reading a query and returning the user documents that the system considers relevant to the information need expressed by this query
    - Binary-retrieval or Ranked-retrieval
  - Peraphs reading user's feedback on the documents retrieved and using it for performing an improved retrieval pass.

Representing textual documents and information needs

- String searching is inappropriate for IR (best suited for data retrieval)
  - Computationally hard
  - The occurrence of string x in a document d in neither necessary nor sufficient condition for relevance of d to an information need about x
- It is thus usual to produce an internal representation (IREP) of the 'meaning' of documents (off-line) and of queries (on-line), and to determine their match at retrieval time

Indexing

Indexing: The process by which IREPs of documents and queries are produced

- It typically generates a set of (possibly weighted) index terms (or features) as the IREPs of a document (or query)
- Underlying assumption: the meaning of this set well approximates the meaning of the document (or query)

Indexing in textual IR

- Index terms in textual IR can be:
  - Words (e.g. classification) automatically extracted
  - Stems of words (e.g. class-) automatically extracted
  - N-grams automatically extracted
  - Noun-phrases (e.g. classification if industrial processes) automatically extracted
  - Words (or noun phrases) from a controlled vocabulary (e.g. categorization)
  - …

Incidence Matrix

The result of the indexing process: the ***incidence matrix.***

	d_1	…	d_i	…	d_m
t_1	w_{11}	…	w_{1i}	…	w_{1m}
…	…	…	…	…	…
t_k	w_{k1}	…	w_{ki}	…	w_{km}
…	…	…	…	…	…
t_n	w_{n1}	…	w_{ni}	…	w_{nm}

N.B. W_{ki} can either be binary or real.

$T = \{t_1,..,t_n\}$ is the **dictionary** of the document base

Indexing - Weights

- Weights can be assigned
  1. **Manually** (typically binary weights are used) by trained human indexers or intermediaries who are familiar with
     - The discipline the documents deal with
     - The indexing technique (e.g. the optimum number of terms for an IREP, the controlled vocabulary,..)
     - The contents of the collection (e.g. topic distribution)
  2. **Automatically** (either binary or real weights are used): by indexing processes based on a statistical analysis of word occurrence in the documents, in the query and in the collection.
- Approach 2. is nowadays the only one left in *text retrieval* (cheaper and more effective).
- Approach 1. and 2. has been used in conjunction until recently because of the difficulty in producing effective automatic indexing techniques for non-textual media.

Indexing - considerations

- The use of the same indexing technique for documents and query alike tends to guarantee a correct matching process
- There is indeterminacy in the indexing process: different indexers (human or automatic) do not produce in general the same IREP for the same document!
- Unlike what happened in reference retrieval systems, the on-line availability of the entire document allows the use of the entire document also for indexing

Evaluation and Experimentation

- The evaluation of an IR technique or system is in general accomplished experimentally (instead of analytically)

  1. Effectiveness
     How well it separates/ranks docs w.r.t. the degree of relevance for the user
  2. Efficiency
     How long it takes for indexing, for searching,.. How expressive is the query language. How large is the doc base
  3. Utility
     Quality w.r.t. costs (of design, development, update, use, etc.) paid by involved parties (developers, users, etc.)
  4. Coverage (e.g. for Web search engines)
     The subset of the available docs that the search engine 'covers' (i.e. indexes, provides access to)

- Criteria 1. and 4. are widely used, 3. is hardly quantifiable.

Effectiveness for Binary Retrieval: Precision and Recall

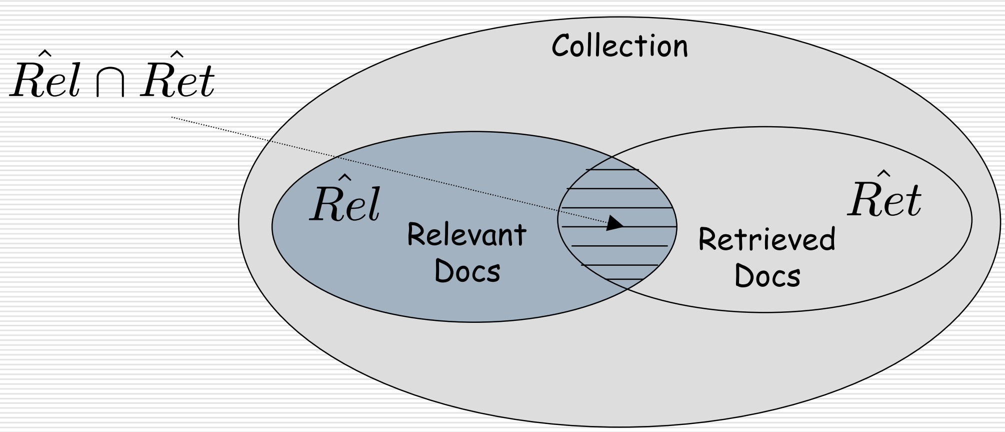
If relevance is assumed to be binary-valued, effectiveness is typically measured as a combination of

- Precision: the "degree of soundness" of the system

$$\pi = Pr(Rel|Ret) = \frac{|\hat{Rel} \cap \hat{Ret}|}{|\hat{Ret}|}$$

- Recall: the "degree of completeness" of the system

$$\rho = Pr(Ret|Rel) = \frac{|\hat{Rel} \cap \hat{Ret}|}{|\hat{Rel}|}$$



Contingency Table

	Relevant	Not Relevant
Retrieved	True positives (**tp**)	False positives (**fp**)
Not Retrieved	False negatives (**fn**)	True negatives (**tn**)

$$\pi = \frac{tp}{tp+fp} \qquad \rho = \frac{tp}{tp+fn}$$

Why NOT using the accuracy $\alpha = \frac{tp+tn}{tp+fp+tn+fn}$?

F measure

- Precision-oriented users
  - Web surfers
- Recall-oriented users
  - Professional searchers, paralegals, intelligence analysts
- A measure that trades-off precision versus recall?
  F-measure (weighted harmonic mean of the precision and recall)

$$F = \frac{(\beta^2+1)\pi\rho}{\beta^2\pi+\rho} \qquad F_{\beta=1} = \frac{2\pi\rho}{\pi+\rho}$$

$\beta < 1$ emphasizes precision!

Evaluation for Ranked retrieval Precision at Recall

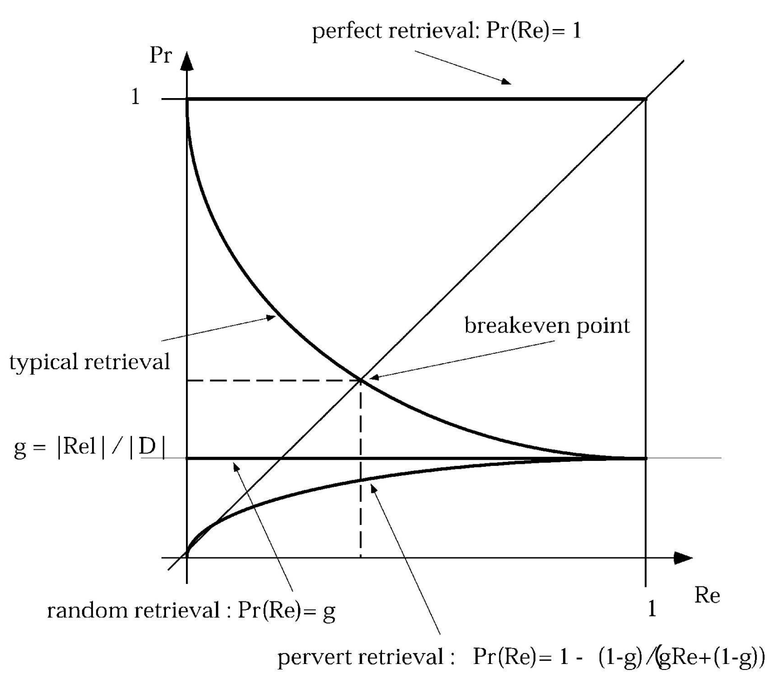
- In a ranked retrieval system, precision and recall are values relative to a rank position r.
- These systems can be evaluated by computing *precision as a function of recall*, i.e. $\pi(\rho)$
  - *What is the precision* $\pi(r(\rho))$ *at the first rank position* $r(\rho)$ *for which recall has a value of* ρ *?*
- We compute this function at each rank position in which a relevant document has been retrieved, and the resulting values are interpolated yielding a *precision/recall plot*
- A unique numerical value of the effectiveness can be obtained by computing e.g. the integral of precision as a function of recall

Collection D, |D| = 100, a query q with |Rel|=20

Rank Pos. q	Rel?	ρ	$\pi(\rho)$
1	Y	1/20=0.05	1/1=1.00
2	Y	2/20=0.10	2/2=1.00
3	N		
4	Y	3/20=0.15	3/4=0.75
5	N		
6	N		
7	Y	4/20=0.20	4/7=0.57
…	…	…	…

Note that

- The effectiveness of a system is typically evaluated by *averaging* over different queries (*macroaveraging*)
  - Different searchers are equally important
  - Partial view of a problem: different methods may work best for different types of queries
- A *typical* precision/recall plot
  - Is monotonically decreasing
  - For $\rho=1$ it takes the value $\pi=g$, where $g=|Rel|/D$ is the *generality* (*frequency*) of the query
- When the document base is big, it is very important to have high precisions for small recall values.
  - Measures such as *precision at 10* (P@10) are often used in place of $\pi(\rho)$
- 'Typical' values are not beyond .4 precision at .4 recall



Precision and Recall Drawbacks

1. The proper estimation of maximum recall requires detailed knowledge of the entire (possibly very large) collection.

2. Precision and recall capture different aspects of the set of retrieved documents. A single measure would be more appropriate.

3. Do not fit the *interactive* retrieval process settings

4. Inadequate for systems which require weak orderings

Benchmark Collections

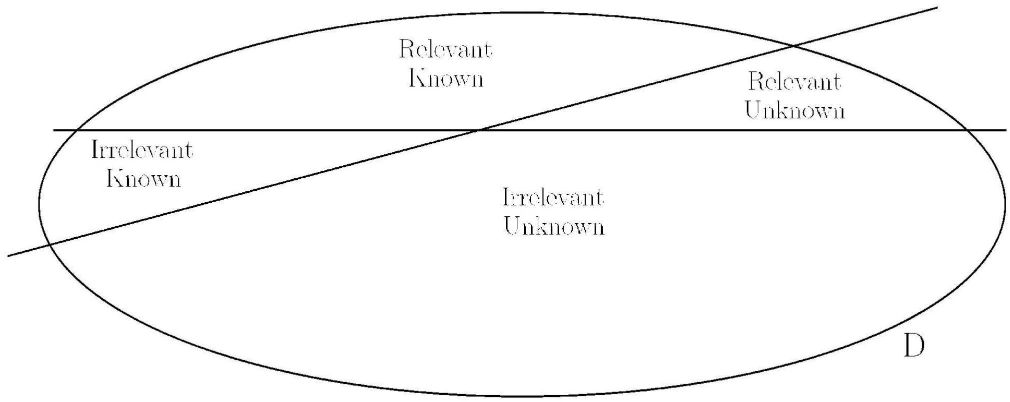
- A set of (up to $O(10^7)$) documents $D = \{d_1,\dots,d_m\}$
- A set of queries (topics in TREC terminology) $Q = \{q_1,\dots,q_l\}$
- A (typically binary) *relevance matrix* of size $m \times l$, who can be either the original query issuers (TREC), or (more often) domain experts.
- A test collection is an abstraction of an operational retrieval environment. It allows to test the relative benefits of different strategies in a controlled way

Limits of the 'scientific' approach

- Two limits
  - R*elevance* judgments tend to be *topicality* judgments
  - It does not consider other aspects such as the user interface and its fitness to the user-seeking behavior
- After all, their construction is the real problem

TREC collection

- TREC web site http://trec.nist.gov
- The corpus: 'Ad hoc' track in the first 8 TREC competitions between '92 and '99.
- Several millions documents and 450 information needs
- In TREC, as in many other big collections relevant documents are identified by a *data-pooling* method thus approximating the set of relevant documents from below



Other Benchmark Collections (used in Text Categorization)

Reuters-21578
- The most widely used in text categorization. It consists of newswire articles which are labeled with some number of topical classifications (zero or more). 9603 train + 3299 test documents

Reuters RCV1
- Newstories, larger than the previous (about 810K documents)

20 Newsgroups
- 18491 articles from the 20 Usenet newsgroups